DOCUMENT RESUME

ED 079 345 TM 002 944

AUTHOR Doppelt, Jerome E.
TITLE How Accurate Is a Test Score?
INSTITUTION Psychological Corp., New York, N.Y.
REPORT NO Bull-50
PUB DATE Jun 56
NOTE 3p.; Reprint from Test Service Bulletin
JOURNAL CIT Test Service Bulletin; n50 p1-3 Jun 1956

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Bulletins; *Measurement Techniques; Reliability; *Scoring Formulas; *Standard Error of Measurement; Statistical Analysis; *Test Results; *True Scores

ABSTRACT

The standard error of measurement as a means for estimating the margin of error that should be allowed for in test scores is discussed. The true score measures the performance that is characteristic of the person tested; the variations, plus and minus, around the true score describe a characteristic of the test. When the standard deviation is used as a measure of the variation of observed scores around the true score, the result is called the standard error of measurement. The standard error of measurement can be used in defining limits around the observed score within which one would be reasonably sure to find the true score. Since, in practice, it is not possible to give a large number of equivalent forms of a test in order to find the characteristic standard error of measurement, it is determined by the reliability coefficient. As measured by the reliability coefficient, reliability means consistency of measurement. It is unfortunately true that a test will have different reliability coefficients depending on the groups of people tested. The standard error of measurement is less subject to this variation. The formula for computing it, which is given, takes into account both the reliability coefficient and the standard deviation for each group. A table is provided of Standard Errors of Measurement for Given Values of Reliability Coefficient and Standard Deviation. (For related document, see TM 002 943, 946.) (DB)

This *Reprint* includes, starting on page 4,
No. 51—APTITUDE, INTELLIGENCE AND ACHIEVEMENT

Test Service Bulletin

No. 50 THE PSYCHOLOGICAL CORPORATION June, 1956

GEORGE K. BENNETT, *President*

Published from time to time in the interest of promoting greater understanding of the principles and techniques of mental measurement and its applications in guidance, personnel work, and clinical psychology, and for announcing new publications of interest. Address communications to 304 East 45th Street, New York, N. Y. 10017.

HAROLD G. SEASHORE, *Editor*
Director of the Test Division

JEROME E. DOPPELT
Assistant Director

DOROTHY M. CLENDENEN
Assistant Director

ALEXANDER G. WESMAN
Associate Director of the Test Division

JAMES H. RICKS, JR.
Assistant Director

ESTHER R. HOLLIS
Advisory Service

HOW ACCURATE IS A TEST SCORE?

EVERY user of test scores knows that no test is perfectly accurate. The score on a test is determined principally by the ability or knowledge of the person who takes it, but the score is also affected by the inaccuracy of the test itself.

It would be helpful if we could know each time we see a score whether it is higher or lower than it should be, and by how much. Unfortunately, no one has ever figured out a practical way to determine the precise amount of error in an individual case. Statistics have been developed, however, for estimating the margin of error we should allow for in test scores. One of the most useful of these is the *standard error of measurement* (SE_M).

At this point, the reader may want to ask, "Doesn't the reliability coefficient tell us how accurate a test is?" The reliability coefficient does, of course, reflect the test's accuracy, but it has two drawbacks: (1) its numerical value depends, to a great extent, on the spread of scores in the group of people tested,* and (2) it does not help us directly in evaluating the scores earned by individual applicants and counselees. The SE_M avoids these two disadvantages. Later in this article, we will show how to compute the SE_M and present a table for estimating it for most tests.

Let us consider a practical situation in which it would be useful to have a measure of the accuracy of a test score. Suppose we have an opening for a junior executive in our company. We have a large number of applicants and among them is Henry Smith. He looks good on most counts, but he has a score of 28 on a test of administrative knowledge. The test norms show that a score of 32 would place an applicant within the upper half of all executive applicants and we desire to make our choice from the upper half. Since Smith looks promising in other ways we begin to wonder about his test placement.

If we could test him again, would he get 28 or some other score? Just what is Smith's *true* score on this test? Before we can make sense in talking about the difference between the *true* score and the *observed* or *obtained* score, we need to specify what we mean by *true* score.

Imagine that we have a very large number of comparable forms of our test. (We need not go into the statistics of comparable forms here; let us simply agree that comparable forms are interchangeable. That is, if we had to choose only one form to measure administrative knowledge, we would be equally happy with any one of the forms.) Now suppose we were able to corner Henry Smith and test him with all our tremendous number of equivalent forms. We would find that our hero does not always get the same score. As the number of forms administered gets larger and larger, we would discover that the distribution of Smith's scores begins to resemble the familiar "normal" curve. In this situation, we can reasonably decide that the average of the large number of scores is characteristic of Smith's performance on our test, and we will call this his *true* score.

At the beginning of the article we pointed out that the score on a test reflects primarily what the person tested brings to the task, but partly error of measurement in the test. The true score measures the perform-

*For an illustration, see Wesman, Alexander G. Reliability and Confidence. *Test Service Bulletin*, No. 44, May, 1952.

ance that is characteristic of the person tested; the variations, plus and minus, around the true score describe a characteristic of the test.

When we use the standard deviation as a measure of the variation of observed scores around the true score, the result is called the *standard error of measurement*. Since this statistic has direct interpretable meaning in relation to the "normal" curve, we are in a position to make this statement:

> If we could know both an individual's exact true score and the SE_M which is characteristic of the test, we would know that about 68% of the scores the individual obtained on the vast number of comparable forms fall within one SE_M of his true score. A band stretching two standard errors above and below his true score would include about 95% of his obtained scores, and within three standard errors of the true score would lie over 99% of his scores on the many forms of the test.

Obviously it is useful to be able to say, putting it a little differently, that for about two thirds of all people tested, the observed scores lie within one SE_M of the true scores — and that for nineteen out of twenty cases the observed score will not be more than two standard errors away from the true score.

As explained in the Note at the end of this article, we must be quite careful how we make statements like the foregoing. It is not correct to say of an individual with a certain *observed* score that the odds are two out of three that his *true* score is within one SE_M of the score he got. But in the practical instance, we can use the SE_M in defining limits around the observed score within which we would be *reasonably sure* to find the true score. Whether the "reasonable limits" (as Professor Gulliksen has called them) will be one, two, or three times the SE_M will depend on the level of confidence the test user desires. The surer he wants to be of not making a mistake in locating the true score, the broader the margin of error he must allow for and therefore the less definite and precise will be the indication given by the test. The broader the score band we allow for each job applicant, for example, the greater the likelihood that his true score will be within it, but the harder it will be to tell the applicants apart.

Coming back to the case of Henry Smith, let us suppose that the test manual reveals that the SE_M is 3 points. If we establish "reasonable limits" of one SE_M on either side of the observed score, the band for Smith would extend over the score range 25-31. And since a score of 32 is needed before a person may be considered as belonging to the top half of executive trainees, we may decide that Smith does not belong in the top half of the group. We are not willing to act as if his true score is 32.

We could have established wider "reasonable limits," say 2 or 3 SE_M on either side of the observed score. We would then have greater confidence that our location of the true score *within the band* is correct. This extra confidence costs us something. We pay for it by having more people to be considered as possibilities. When there are many applicants, we usually want to reduce the number of eligible candidates even though we increase the possibility of making a wrong decision about the true score of some of them.

Since in practice we cannot give a large number of equivalent forms of a test in order to find the characteristic standard error of measurement, how *do* we determine it? The answer to this takes us back to the *reliability coefficient*.

As measured by the reliability coefficient, reliability means consistency of measurement. If the individuals of a group remain in about the same relative positions or ranks after successive testings, the test is "reliable" *for that group*. It is unfortunately true that a test will have different reliability coefficients depending on the groups of people tested: higher coefficients for groups with a wide spread of scores and lower ones for groups with scores bunched more closely together.

The SE_M is less subject to this variation; the formula for computing it takes into account both the reliability coefficient *and* the standard deviation for each group. The formula is simple:

$$SE_M = SD\sqrt{1 - r_{tt}}$$

where SD is the standard deviation of the obtained scores of a group and r_{tt} is the reliability coefficient computed for the same group.*

Like a true score for an individual, the SE_M for a test should be just one definite number if it is really a characteristic of the test rather than of the people tested. But if we look in a test manual, we may see that there appear to be differences among standard errors of measurement computed for different groups. For example, the SE_M is reported for each of nine groups on the Numerical Test in the *Personnel Tests for Industry* series. The values range from 1.7 to 2.4. The explanation is that we have no way of computing the exact value of the SE_M — the formula merely provides an *estimate* of the SE_M. Estimates, of course, can be expected to differ. In any situation where we cannot obtain the true value of a statistic, it is advisable to have as many es-

*We cannot automatically say that the more accurate or reliable of two tests is the one which has the lower value for its SE_M. As may be seen from the computing formula, the SE_M is tied in with the score units in which the standard deviation is expressed. A test with a standard deviation of 16 points may have the same reliability as a test with a standard deviation of 8 points. However the SE_M of the first test will be numerically twice that of the second.

Standard Errors of Measurement for Given Values of Reliability Coefficient and Standard Deviation

SD	Reliability Coefficient					
	.95	.90	.85	.80	.75	.70
30	6.7	9.5	11.6	13.4	15.0	16.4
28	6.3	8.9	10.8	12.5	14.0	15.3
26	5.8	8.2	10.1	11.6	13.0	14.2
24	5.4	7.6	9.3	10.7	12.0	13.1
22	4.9	7.0	8.5	9.8	11.0	12.0
20	4.5	6.3	7.7	8.9	10.0	11.0
18	4.0	5.7	7.0	8.0	9.0	9.9
16	3.6	5.1	6.2	7.2	8.0	8.8
14	3.1	4.4	5.4	6.3	7.0	7.7
12	2.7	3.8	4.6	5.4	6.0	6.6
10	2.2	3.2	3.9	4.5	5.0	5.5
8	1.8	2.5	3.1	3.6	4.0	4.4
6	1.3	1.9	2.3	2.7	3.0	3.3
4	.9	1.3	1.5	1.8	2.0	2.2
2	.4	.6	.8	.9	1.0	1.1

This table is based on the formula $SE_M = SD\sqrt{1 - r_{tt}}$. For most purposes the result will be sufficiently accurate if the table is entered with the reliability and standard deviation values nearest those given in the test manual. Be sure the standard deviation and the reliability coefficient are for the same group of people

timates of that value as practical. In the case of PTI-Numerical, we can be comfortable with the conclusion that the SE_M is about 2 points.

Many test manuals give both reliability coefficients and standard errors of measurement for the convenience of the user. When the SE_M is not given, it can be estimated readily by use of the reliability coefficient, provided the manual also states the standard deviation of the particular group of people on which the reliability coefficient is based. It is well worth the test user's time to make this computation; the table at the left permits an approximation to be made easily without any figuring.

If, as is too often the case, the manual does not present the standard deviation of the group for which the reliability coefficient is reported, it would be advisable for the user to write a letter to the test author.—J. E. D.

NOTE: As textbooks usually point out, it is correct to make a statement of probability (such as "68% of the scores" or "two out of three times") *only* when the SE_M is applied to the *true* score. If a test has a standard error of 5.5, it is not correct to say of a person who obtains a score of 48 that the chances are two out of three that his true score is between 42.5 and 53.5. This person's true score is a definite number, although we do not know what it is. The statement that his true score lies between 42.5 and 53.5 is either true or false. Intermediate probabilities like "two out of three" or "one out of twenty" cannot properly be attached to it. The "reasonable limits" idea simply helps us to avoid making a mathematical statement of probability which would be technically inaccurate. Precise statements of probability in relation to confidence intervals are possible but lie outside the scope of this article.

Readers who want to pursue this and other fine points regarding the standard error of measurement will find good treatments in, among others, the following texts:

H. Gulliksen. *Theory of mental tests.* New York: Wiley, 1950.

T. L. Kelley. *Fundamentals of statistics.* Cambridge: Harvard University Press, 1947.

E. F. Lindquist. *A first course in statistics.* Boston: Houghton Mifflin, 1942.

A Book of BASIC READINGS ON THE MMPI

This book, edited by G. S. Welsh and W. G. Dahlstrom, brings together in one place 66 of the most important articles on the *Minnesota Multiphasic Personality Inventory* that have appeared in its fifteen years of steadily widening use. More than 600 additional articles are listed in the bibliography, plus nearly 200 supplementary references.

The articles are grouped in ten sections: Theory, Construction, Coding, New Scales, Profile Analysis, Diagnostic Profiles, Psychiatric Problems, Medical Problems, Therapy, and General Personality. xviii + 656 pages,